JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS 113, 382–389 (1986)

Finite-State Approximations for Denumerable Multidimensional State Discounted Markov Decision Processes*

ONÉSIMO HERNÁNDEZ-LERMA

Departamento de Matemáticas, Centro de Investigación del I.P.N., Apartado Postal 14-740, 07000 México, D.F., Mexico

Submitted by E. Stanley Lee

A finite-state iterative scheme introduced by White (*in* "Recent Developments in Markov Decision Processes" (R. Hartley, L. C. Thomas, and D. J. White, Eds.), Academic Press, New York, 1980) to approximate the value function of denumerable-state Markov decision processes is extended to denumerable *multidimensional* state space. Under essentially the same assumptions given in (op cit.), a simpler proof of the convergence theorem is obtained together with convergence rates. The iterative scheme is used to determine an asymptotically discount optimal policy, which in turn can be used to obtain a discount optimal stationary policy.

1. INTRODUCTION

In a series of papers [10, 11, 12], White has considered several finite-state approximation schemes for the optimal value function of discounted Markov decision processes (MDP's), the origin of which can be traced back to Fox [2]. In particular, for a given decision model (I, A, p, r, β), where $I = \{1, 2, \ldots\}$ is a denumerable set of states; A is the action space; $r(i, a)$ is the reward function defined on $K := \{(i, a)\colon i \in I,\ a \in A(i)\}$, with $A(i) \subset A$ the set of admissible actions in state i; $p_{ij}(a)$, $a \in A(i)$, are the transition probabilities; and β is the discount rate, $0 \leqslant \beta < 1$, White introduced in [10] the following method of successive approximations: Define a sequence of functions $\{w_n(\cdot)\}$ by

$$w_0(i) := u(i), \qquad i \in I,$$

and for $n = 1, 2, \ldots$,

$$\begin{aligned} w_n(i) &:= \sup_{a \in A(i)} \left\{ r(i, a) + \beta \sum_{j \in I} p_{ij}(a)\, w_{n-1}(j) \right\} && \text{if} \quad i \leqslant n, \\ &:= u(i) && \text{if} \quad i > n, \end{aligned} \tag{1}$$

* This research was supported in part by the Consejo Nacional de Ciencia y Tecnologia under Grant PCCBBNA 020630.

0022-247X/86 $3.00

where $u(\cdot)$ is a given bounded function on I. He then showed [10, Theorem 3] that, for a bounded reward function, if

$$\sup_{(i,a)\in K} \sum_{j \geq N+i} p_{ij}(a) \to 0 \qquad \text{as} \quad N \to \infty, \tag{2}$$

then $\{w_n(\cdot)\}$ converges pointwise to the unique bounded solution, say $w^*(\cdot)$, of the equation

$$w^*(i) = \sup_{a \in A(i)} \left\{ r(i,a) + \beta \sum_{j \in I} p_{ij}(a)\, w^*(j) \right\}. \tag{3}$$

This result was extended in [12] to unbounded rewards.

There are many situations, however, like, e.g., in the control of priority queueing systems [5] or computer communication networks [7], where the "state" of the system is multidimensional and an iterative scheme like White's (1), in a proper setting, would be very useful. Thus in the present paper we consider an MDP (S, A, p, r, β), where A, p, r, and β are defined as above, but now the state space S is a denumerable *multidimensional* set, say $S \subset \mathbb{Z}^d$, the space of d-dimensional vectors with integer components, and instead of White's scheme (1), we consider functions $v_n(\cdot)$ defined by

$$v_0(x) := 0, \qquad x \in S,$$

and for $n = 1, 2, \ldots,$

$$\begin{aligned} v_n(x) &:= \sup_{a \in A(x)} \left\{ r(x,a) + \beta \sum_{\|y\| < n} p_{xy}(a)\, v_{n-1}(y) \right\} && \text{if} \quad \|x\| \leqslant n, \\ &:= 0 && \text{if} \quad \|x\| > n, \end{aligned} \tag{4}$$

where $\|x\| = \sup\{|x_i| : i = 1, \ldots, d\}$. Note that (1) is reduced to a form analogous to (4) simply by taking $u(\cdot) = 0$. On the other hand, we could also introduce in (4) a function $u(\cdot) \neq 0$ as in (1), but by doing so nothing is really gained except complicating the notation; we have thus prefered the simpler scheme (4).

Our first result (Theorem 1 below) is a slight generalization of White's convergence Theorem 3 in [10] in that, under essentially the same condition as White's (2) above, we give precise bounds for the convergence, uniform on bounded sets, of $\{v_n(\cdot)\}$ to the optimal value function. In the proof of Theorem 1 we make heavy use of the similarity between the iterative schemes (1) and (4), and the nonstationary value-iteration (NVI) scheme introduced by Federgruen and Schweitzer [1] for MDP's with finite state and action spaces, and extended by the present author [3] to

denumerable state space semi-MDP's. The NVI scheme has been used to determine *adaptive* policies for Markov [4] and semi-Markov [3] decision processes depending on unknown parameters. An approach similar to that in [3, 4] is used below (Sect. 3) to determine an asymptotically discount optimal policy based on (4). Henceforth we shall refer to the iterative scheme (4) as a *truncated NVI* scheme.

In Section 2, convergence of the functions $v_n(\cdot)$ in (4) is proved, and in Section 3 an asymptotically discount optimal policy is defined.

2. The Truncated NVI Scheme

Let (S, A, p, r, β) be the decision model introduced in Section 1, where $S \subset \mathbb{Z}^d$ is a denumerable set, and A, the action (or control) set is a metric space endowed with the Borel σ-algebra. Throughout the following we assume: there is a constant R such that

(A1) $|r(x, a)| \leqslant R$ for all $(x, a) \in K$.

The case of unbounded rewards can be treated as in [12].

For each $n = 0, 1, \ldots$, let X_n and A_n be the state and action at the nth stage, respectively. As usual [3–6, 8, 9], *a* (nonrandomized) *policy* is defined as a sequence $D = (D_n, n = 0, 1, \ldots)$ of measurable functions such that for each n, D_n specifies which action to choose at the nth decision epoch given the current state X_n and the sequence X_k, A_k, $k = 0, 1, \ldots, n-1$, of previous states and actions. We shall denote by $\mathscr{D}$ the set of all policies. $D \in \mathscr{D}$ is said to be a *memoryless* or *Markov* policy if, for each n, D_n depends only on X_n. Thus for a Markov policy $D = (D_n)$, $D_n \in F$ for all $n = 0, 1, \ldots$, where F is the set of all functions $f: S \to A$ such that $f(x) \in A(x)$, $x \in S$. A Markov policy $D = (D_n)$ such that D_n is the same function, say $D_n = f$, for all n is called *stationary* and in this case we write $D = (f, f, \ldots) \in F$, or simply, $f \in F$.

Now let

$$V(D, x) := E_x^D \left\{ \sum_{n=0}^{\infty} \beta^n r(X_n, A_n) \right\}, \qquad D \in \mathscr{D}, \quad x \in S,$$

be the expected total discounted reward when policy D is employed and the initial state is x, and let

$$v^*(x) := \sup_{D \in \mathscr{D}} V(D, x), \qquad x \in S.$$

It is well known (see, e.g., [6, Sect. 6]) that $v^*(\cdot)$ is the unique bounded solution of the *optimality equation*

$$v^*(x) = \sup_{a \in A(x)} \left\{ r(x, a) + \beta \sum_y p_{xy}(a)\, v^*(y) \right\}, \qquad x \in S. \tag{5}$$

To prove convergence of the truncated NVI scheme (4) we shall impose the following condition (cf. (2)):

(A2) $\varepsilon(n) := \sup_{(x,a) \in K} \sum_{\|y\| > n} p_{xy}(a) \to 0$ as $n \to \infty$.

Note that $\varepsilon(n) \geqslant \varepsilon(n+1)$ for all n. Let us define

$$\|u\|_n := \sup_{\|x\| \leqslant n} |u(x)|,$$

for any function u on S, and $n = 0, 1, \ldots$.

THEOREM 1. *Under assumptions* (A1, A2), *the sequence* $\{v_n(\cdot)\}$ *converges to* $v^*(\cdot)$ *uniformly on bounded sets. More precisely, there exists a constant* C *such that, for all* n,

$$\|v_n - v^*\|_n \leqslant C \cdot \max\{\beta^{[n/2]}, \varepsilon([n/2])\}. \tag{6}$$

Proof. (cf. [1, Theorem 3.1]). We shall use the following result (see, e.g., Lemma 3.3 in [6, p. 17]): If u and v are real-valued bounded functions on an arbitrary space, then

$$|\sup_x u(x) - \sup_x v(x)| \leqslant \sup_x |u(x) - v(x)|. \tag{7}$$

Now, under (A1), is clear that $v^*(\cdot)$ and $v_n(\cdot)$ are bounded: for all n,

$$\|v^*\| \leqslant C_1 \qquad \text{and} \qquad \|v_n\| \leqslant R \sum_{k=0}^{n} \beta^k \leqslant C_1,$$

where $C_1 = R/(1-\beta)$. Therefore, from (4), (5), and (7), we obtain, for $\|x\| \leqslant n+1$,

$$\begin{aligned} |v_{n+1}(x) - v^*(x)| &\leqslant \sup_{a \in A(x)} \left| \beta \sum_y p_{xy}(a)[v_n(y) - v^*(y)] \right| \\ &\leqslant \beta \sup_{a \in A(x)} \left\{ \sum_{\|y\| \leqslant n} p_{xy}(a)\, |v_n(y) - v^*(y)| + \sum_{\|y\| > n} p_{xy}(a)\, |v^*(y)| \right\}, \end{aligned}$$

so that

$$\|v_{n+1} - v^*\|_{n+1} \leqslant \beta\, \|v_n - v^*\|_n + \beta\, \|v^*\|\, \varepsilon(n).$$

Therefore,

$$\begin{aligned}\|v_{n+m} - v^*\|_{n+m} &\leqslant \beta^m \|v_n - v^*\|_n + \|v^*\| \sum_{k=1}^{m} \beta^k \varepsilon(n+m-k) \\ &\leqslant \beta^m \|v_n - v^*\| + \varepsilon(n) \|v^*\| \sum_{k=1}^{m} \beta^k \\ &\leqslant 2C_1 \beta^m + \varepsilon(n) C_1 \beta/(1-\beta),\end{aligned}$$

i.e.,

$$\|v_{n+m} - v^*\|_{n+m} \leqslant C \cdot \max\{\beta^m, \varepsilon(n)\}, \tag{8}$$

where $C = 2 \cdot \max\{2C_1, C_1\beta/(1-\beta)\}$. Finally making the substitution $k = n+m$, with $n = [k/2]$ and $m = k - [k/2] \geqslant [k/2]$, inequality (8) reduces to (6), which completes the proof of the theorem. ∎

As noted already in Section 1, the proof of Theorem 1 exploits the *similarity* between the iterative scheme (4) and the NVI scheme of Federgruen and Schweitzer [1, Theorem 3.1]. Note, however, that neither the results in [1] for MDP's with finite state and action spaces, nor the results in [3] for semi-MDP's with denumerable state space are applicable to the truncated NVI scheme (4). In other words, (4) is similar to, but it is *not* a special case of the NVI schemes in [1, 3, 4].

Last, we should mention that inequality (6) holds for the "difference" between $v_{n+1}(\cdot)$ and $v_n(\cdot)$; namely, for all n,

$$\|v_{n+1} - v_n\|_n \leqslant C \cdot \max\{\beta^{[n/2]}, \varepsilon([n/2])\}.$$

This can be obtained as (6).

3. Asymptotically Discount Optimal Policies

In many applications it is desirable to determine policies which will give the optimal value function, and this can be done in a number of ways [1, 10]. Here we will determine such policies using the iterative scheme (4) and the function $\phi: K \to \mathbb{R}$ defined by

$$\phi(x, a) = r(x, a) + \beta \sum_{y \in S} p_{xy}(a)\, v^*(y) - v^*(x). \tag{9}$$

This function is commonly used [1, 3–5, 8, 9] as a measure of the difference (or discrepancy [8]) between an optimal action in state x and any action $a \in A(x)$. In particular, the optimality equation (5) and the optimality criterion [6, 9] can be written as follows.

LEMMA 1. (a) $\sup_{a \in A(x)} \phi(x, a) = 0$, $x \in S$.

(b) *A stationary policy* $g \in F$ *is discount optimal if and only if* $\phi(x, g(x)) = 0$ *for all* $x \in S$.

In Lemma 1(b), recall that a policy $D \in \mathscr{D}$ is said to be *discount optimal* if

$$V(D, x) = v^*(x), \qquad x \in S. \tag{10}$$

Using ϕ we can now define asymptotic discount optimality.

DEFINITION 1. A Markov policy $\{f_n(\cdot)\}$ is *asymptotically discount optimal* (ADO) if, for every $x \in S$, $\phi(x, f_n(x)) \to 0$ as $n \to \infty$.

Comments. Asymptotic discount optimality is related to the following concept due to Schäl [9]: A policy $D \in \mathscr{D}$ is asymptotically discount optimal in the sense of Schäl (ADOS) if, for every $x \in S$,

$$V_N(D, x) - E_x^D v^*(X_N) \to 0 \qquad \text{as} \quad N \to \infty, \tag{11}$$

where

$$V_N(D, x) := E_x^D \left\{ \sum_{n=N}^{\infty} \beta^{n-N} r(X_n, A_n) \right\},$$

is the expected total reward from stage N onwards discounted at stage N. As noted by Schäl [9] every discount optimal policy is ADOS, since by Bellman's principle of optimality [6, p. 109], Eq. (10) implies $V_N(D, x) = E_x^D v^*(X_N)$ for all N. On the other hand, the relation between ADO (Definition 1) and ADOS is obtained from the fact [9, Theorem 4.12] (see also [3, 4]) that the left-hand side of (11) can be written as

$$V_N(D, x) - E_x^D v^*(X_N) = E_x^D \left\{ \sum_{n=N}^{\infty} \beta^{n-N} \phi(X_n, A_n) \right\}.$$

Finally ADO and discount optimality are related as follows.

THEOREM 2. *In addition to* (A1) *and* (A2), *let us assume*:

(A3) *For all* $x \in S$, $A(x)$ *is compact, and,*

(A4) *for all* $x, y \in S$, *the functions* $a \to r(x, a)$ *and* $a \to p_{xy}(a)$ *are continuous on* $A(x)$.

Then, if $\{f_n(\cdot)\}$ *is an ADO Markov policy, there exists a subsequence* $\{f_{n'}(\cdot)\}$ *and a discount optimal stationary policy* $g(\cdot) \in F$ *such that* $\{f_{n'}(\cdot)\}$ *converges pointwise to* $g(\cdot)$.

Proof. First note that, since $v^*(\cdot)$ is bounded, (A4) implies that $a \to \phi(x, a)$ is continuous on $A(x)$, $x \in S$ [6, 9]. Now, for each x, $\{f_n(x)\}$ is a sequence in the compact set $A(x)$ and therefore, there is a subsequence which (to simplify notation) we also denote by $\{f_n(x)\}$ and a point $g(x) \in A(x)$ such that $g(x) = \lim f_n(x)$. This can be done for each $x \in S$, and since every subsequence of an ADO policy is also ADO, we see that $\phi(x, g(x)) = \lim \phi(x, f_n(x)) = 0$; that is, $g \in F$ is an optimal stationary policy. ∎

Thus Theorem 2 provides a way to approximate an optimal stationary policy using an ADO policy. In turn, using (4) we can determine an ADO policy as follows, assuming (which we do) that (A1)–(A4) hold. Let $\{f_n(\cdot)\}$ be a sequence of functions such that $f_0(\cdot) \in F$ is arbitrary, and for $n \geqslant 1$, we define

$$\begin{aligned} f_n(x) &:= \arg \max_{a \in A(x)} \left\{ r(x, a) + \beta \sum_{\|y\| < n} p_{xy}(a)\, v_{n-1}(y) \right\} && \text{if } \|x\| \leqslant n, \\ &:= \text{arbitrary point in } A(x) && \text{if } \|x\| > n. \end{aligned} \tag{12}$$

In other words, $f_n(\cdot)$ is just the maximizer of the right-hand side of (4). Since $f_n(\cdot) \in F$, $D^* = \{f_n(\cdot)\}$ is a Markov policy.

THEOREM 3. *Assuming* (A1)–(A4), D^* *is an ADO policy; in fact, as* $n \to \infty$,

$$\|\phi\|_n := \sup_{\|x\| \leqslant n} |\phi(x, f_n(x))| \to 0. \tag{13}$$

Proof. For $\|x\| \leqslant n$, we have

$$\begin{aligned} \phi(x, f_n(x)) &= \phi(x, f_n(x)) - v_n(x) + v_n(x), \\ &= \beta \sum_{\|y\| < n} p_{xy}(f_n(x))[v^*(y) - v_{n-1}(y)] \\ &\quad + \beta \sum_{\|y\| \geqslant n} p_{xy}(f_n(x))\, v^*(y) + [v_n(x) - v^*(x)], \end{aligned}$$

and therefore,

$$\|\phi\|_n \leqslant \beta \|v_{n-1} - v^*\|_{n-1} + \beta \|v^*\|\, \varepsilon(n-1) + \|v_n - v^*\|_n.$$

Thus (13) follows from Theorem 1 and (A2). ∎

Combining Theorems 2 and 3, a discount optimal stationary policy can be determined using the functions $f_n(\cdot)$ in (12).

REFERENCES

1. A. FEDERGRUEN AND P. J. SCHWEITZER, Nonstationary Markov decision problems with converging parameters, *J. Optim. Theory Appl.* **34** (1981), 207–241.
2. B. L. FOX, Finite-state approximations to denumerable-state dynamic programs, *J. Math. Anal. Appl.* **34** (1971), 665–670.
3. O. HERNÁNDEZ-LERMA, Nonstationary value-iteration and adaptive control of discounted semi-Markov processes, *J. Math. Anal. Appl.* **112** (1985), 435–445.
4. O. HERNÁNDEZ-LERMA AND S. I. MARCUS, Adaptive control of discounted Markov decision chains, *J. Optim. Theory Appl.* **46** (1985), 227–235.
5. O. HERNÁNDEZ-LERMA AND S. I. MARCUS, Optimal adaptive control of priority assignment in queueing systems, *Systems Control Lett.*, **4** (1984), 65–72.
6. K. HINDERER, Foundations of non-stationary dynamic programming with discrete time parameter, Lecture Notes Oper. Res. **33** (1970).
7. H. KOBAYASHI AND A. G. KONHEIM, Queueing models for computer communications system analysis, *IEEE Trans. Commun.* **COM-25** (1977), 2–29.
8. P. MANDL, Estimation and control in Markov chains, *Adv. Appl. Probab.* **6** (1974), 40–60.
9. M. SCHÄL, Estimation and Control in Discounted Stochastic Dynamic Programming, Preprint No. 428, Inst. Angew. Math., Univ. of Bonn, 1981.
10. D. J. WHITE, Finite state approximations for denumerable state infinite horizon discounted Markov decision processes: The method of successive approximations, *in* "Recent Developments in Markov Decision Processes" (R. Hartley, L. C. Thomas, and D. J. White, Eds.) Academic Press, New York, 1980.
11. D. J. WHITE, Finite state approximations for denumerable state infinite horizon discounted Markov decision processes, *J. Math. Anal. Appl.* **74** (1980), 292–295.
12. D. J. WHITE, Finite state approximations for denumerable state infinite horizon discounted Markov decision processes with unbounded rewards, *J. Math. Anal. Appl.* **86** (1982), 292–306.